Coordination Centre

DARWIN EU® - A transportable framework for evaluating pregnancy episode algorithms in heterogeneous real-world databases

Aniek F. Markus, Julieta Politi, Anum Zahra, Maarten van Kessel, Natasha Yefimenko, Adam Black, Ger Inberg, Berta Raventós, Maxim Moinat, Daniel R. Morales, María-Clara Restrepo-Méndez, Katia Verhamme, Talita Duarte-Salles

OHDSI Community Call – November 17th 2026

Enabling pregnancy and mother-child research in the DARWIN EU® Data Network

1. To **identify data partners** in the DARWIN EU® Data Network with available data on pregnancy, pregnancy start and end dates, pregnancy outcomes and mother and child linkage.
2. To **optimise and validate OMOP pregnancy algorithms** for DARWIN EU® databases with pregnancy-related data.
3. To **establish a convention for the mapping** of pregnancy-related data and mother and child linkage data to the OMOP CDM, exploring the use of the Perinatal Extension Table (PET).
4. To **characterise pregnancies and mother-child linkage** over time for DARWIN EU® databases with pregnancy-related data.
5. To **create a database with aggregate counts** related to pregnancy and mother-child linkage to support feasibility assessment.

Classified as confidential by the European Medicines Agency

Goal



To develop an **accurate and generalisable** pregnancy episode inference algorithm



Algorithm = set of rules that combines clinical events and temporal logic to derive **pregnancy start, end & outcome**



Requires **robust performance** across **diverse data sources**

Classified as confidential by the European Medicines Agency

Existing pregnancy episode inference algorithms



JAMIA Open, 2023, **6(3)**, ooad067
https://doi.org/10.1093/jamiaopen/ooad067
Advance access publication 9 August 2023
Research and Applications

Research and Applications

Who is pregnant? Defining real-world data-based pregnancy episodes in the National COVID Cohort Collaborative (N3C)

Sara E. Jones, MS[1], Katie R. Bradwell, PhD[2], Lauren E. Chan, PhD, RD[3], Julie A. McMurry, MPH[4], Courtney Olson-Chen, MD[5], Jessica Tarleton, MD[6], Kenneth J. Wilkins, PhD[7], Victoria Ly, MD[5], Saad Ljazouli, MS[2], Qiuyuan Qin, MS[8], Emily Groene Faherty, PhD[9], Yan Kwan Lau, PhD[10], Catherine Xie[8], Yu-Han Kao, PhD[10], Michael N. Liebman, PhD[11], Federico Mariona, MD[12,13], Anup P. Challa, MSEng, EIT[14], Li Li, MD[10], Sarah J. Ratcliffe, PhD[15], Melissa A. Haendel, PhD[3], Rena C. Patel, MD, MPH, MPhil[16], Elaine L. Hill, PhD*[,5,8]; on behalf of the N3C Consortium

[1]Office of Data Science and Emerging Technologies, Nati
Rockville, MD 20852, United States, [2]Palantir Technologie
Oregon State University, Corvallis, OR 97331, United State
Medical Campus, Aurora, CO 80045, United States, [5]Depa
Rochester, NY 14620, United States, [6]Department of Obste
United States, [7]Biostatistics Program, Office of the Direct
Institutes of Health, Bethesda, MD 20892, United States, [8]
Rochester, NY 14618, United States, [9]School of Public Hea
Stamford, CT 06902, United States, [11]IPQ Analytics, LLC, K
United States, [13]Wayne State University, Detroit, MI 48202
Vanderbilt University, Nashville, TN 37212, United States,
22903, United States, and [16]Department of Medicine and C

***Corresponding author:** Elaine L. Hill, PhD, Department of Public He
States (elaine_hill@urmc.rochester.edu)

Author Contributions: Sara E. Jones and Dr Katie R. Bradwell cont
L. Hill contributed equally and are considered as senior authors

Journal of the American Medical Informatics Association, 2024, **31(12)**, 2789–2799
https://doi.org/10.1093/jamia/ocae195
Advance access publication 24 July 2024
Research and Applications

Research and Applications

Pregnancy episodes in *All of Us*: harnessing multi-source data for pregnancy-related research

Louisa H. Smith, PhD[1,2,*], Wanjiang Wang, MS[1], Brianna Keefe-Oates, PhD[1,2]

[1]Department of Public Health and Health Sciences, Bouvé College of Health Sciences, Northeastern University, Boston, MA 02115, United States, [2]Roux Institute, Northeastern University, Portland, ME 04101, United States

*Corresponding author: Louisa H. Smith, PhD, Roux Institute, Northeastern University, 100 Fore St, Portland, ME 04101, United States (l.smith@northeastern.edu)

Abstract

Objectives: The National Institutes of Health's *All of Us* Research Program addresses gaps in biomedical research by collecting health data from diverse populations. Pregnant individuals have historically been underrepresented in biomedical research, and pregnancy-related research is often limited by data availability, sample size, and inadequate representation of the diversity of pregnant people. *All of Us* integrates a wealth of health-related data, providing a unique opportunity to conduct comprehensive pregnancy-related research. We aimed to identify pregnancy episodes with high-quality electronic health record (EHR) data in *All of Us* Research Program data and evaluate the program's utility for pregnancy-related research.

Materials and Methods: We used a previously published algorithm to identify pregnancy episodes in *All of Us* EHR data. We described these pregnancies, validated them with *All of Us* survey data, and compared them to national statistics.

Results: Our study identified 18 970 pregnancy episodes from 14 234 participants; other possible pregnancy episodes had low-quality or insufficient data. Validation against people who reported a current pregnancy on an *All of Us* survey found low false positive and negative rates. Demographics were similar in some respects to national data; however, Asian-Americans were underrepresented, and older, highly educated pregnant people were overrepresented.

Discussion: Our approach demonstrates the capacity of *All of Us* to support pregnancy research and reveals the diversity of the pregnancy cohort. However, we noted an underrepresentation among some demographics. Other limitations include measurement error in gestational age and limited data on non-live births.

Conclusion: The wide variety of data in the *All of Us* program, encompassing EHR, survey, genomic, and fitness tracker data, offers a valuable resource for studying pregnancy, yet care must be taken to avoid biases.

Key words: pregnancy; pregnancy outcomes; gestational age; electronic health records; real-world data.

- **HIPPS algorithm:**

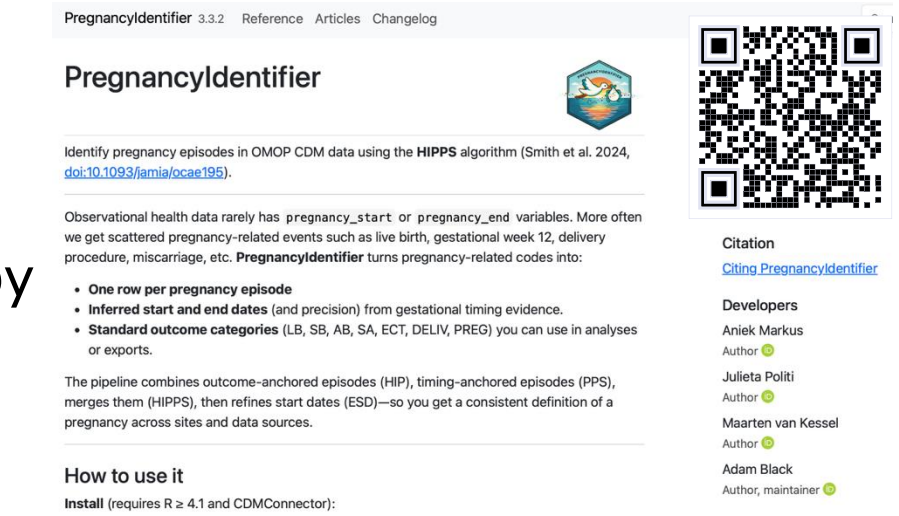
  (1) an extension of Matcho et al. (2018)

  (2) adds a step to detect gestational age-specific signatures of a progressing pregnancy for further episode support

  (3) pregnancy start date inference

Classified as confidential by the European Medicines Agency

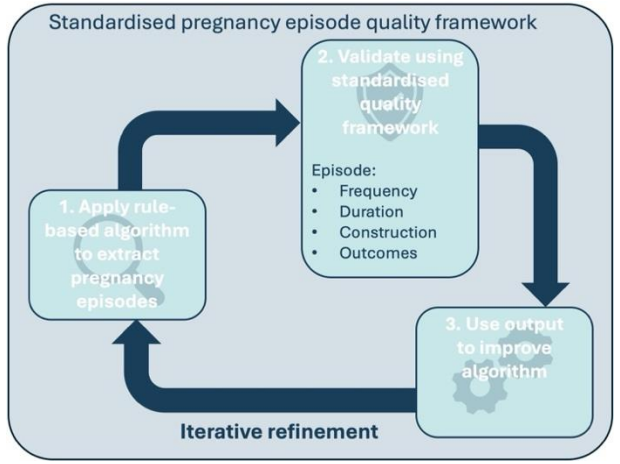
PregnancyIdentifier package



An optimized implementation of the HIPPS algorithm by Jones et al. (2023) to identify:

to identify:

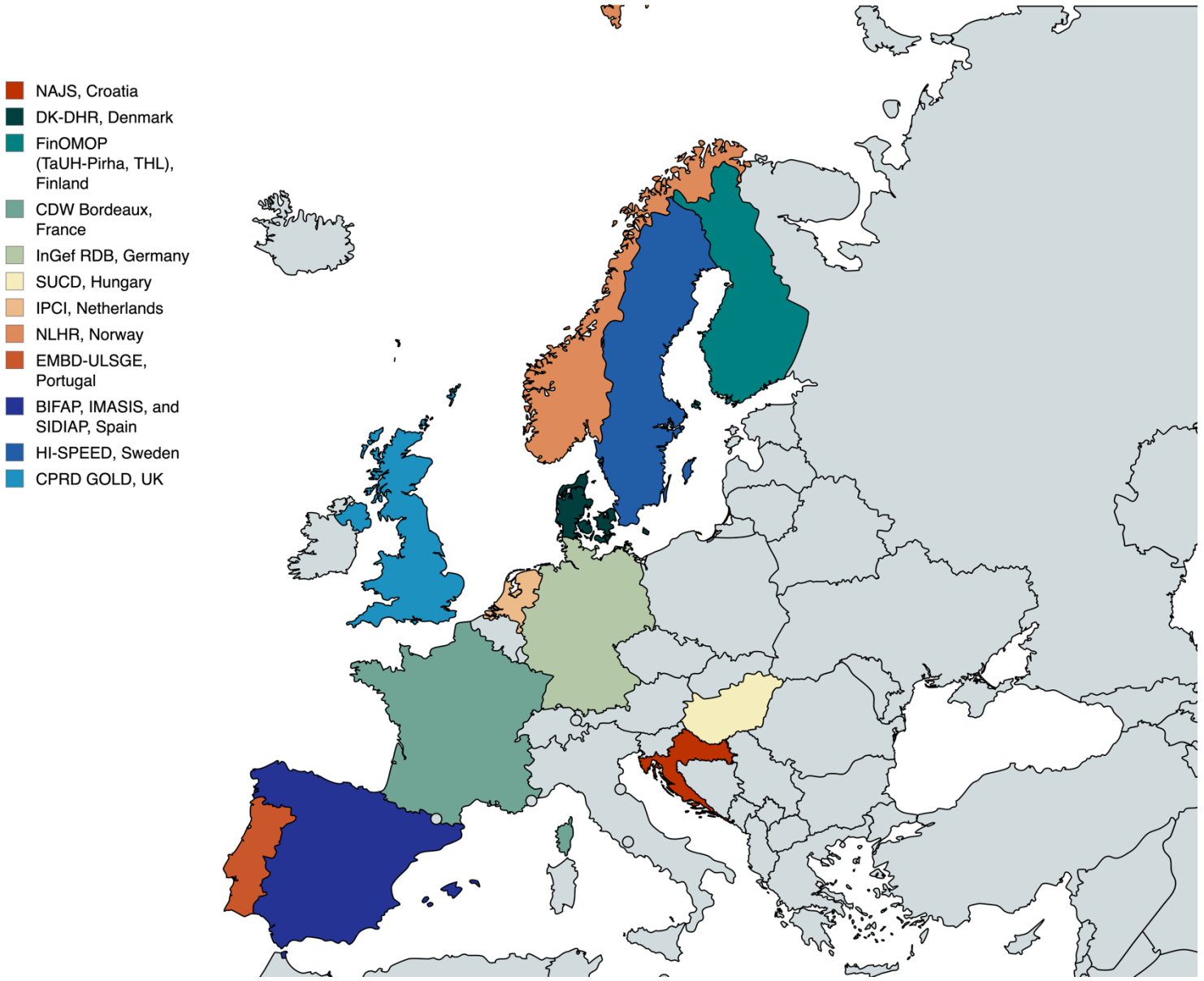
- Pregnancy start date
- Pregnancy end date
- Pregnancy episode outcome (live birth, stillbirth, abortion, spontaneous abortion, ectopic pregnancy, delivery, and pregnancy with unknown outcome)



https://github.com/darwin-eu/pregnancyidentifier

Classified as confidential by the European Medicines Agency

Participating Data Sources



Classified as confidential by the European Medicines Agency

Validation steps

Episode-level

Are individual pregnancy episodes correctly identified?

- Quality evaluation framework covering clinical plausibility, internal consistency, and clinical completeness
- Concordance to locally derived episodes from Perinatal Expansion Table (PET) (where available)

Population-level

Do overall characteristics match known distributions?

- Compare aggregated outputs to national statistics (e.g. birth rates, maternal age at delivery)

Classified as confidential by the European Medicines Agency

Quality evaluation framework

Dimension	Primary metric	Purpose	Secondary metrics	Purpose
Episode frequency	Pregnancy incidence rate per 1,000 person-years recorded as female (at birth) aged 12-55	Does the number of identified pregnancies look plausible?	Number of pregnancy episodes; number of pregnant individuals; maternal age distribution	Do the absolute counts and demographic profile of the pregnant population look plausible?
Episode duration	Gestational duration distribution within the plausible range (0-308 days)	Are episode lengths clinically plausible, without implausible durations or extreme outliers?	Distribution of gestational age at anchor event (e.g. delivery, first antenatal visit, ultrasound after first trimester, gestational diabetes mellitus); gestational week bins; seasonal and temporal patterns	Do temporal patterns align with expected pregnancy chronology?
Episode construction	Proportion of overlapping pregnancy episodes among individuals with multiple episodes	Are episodes constructed cleanly, without logically incompatible overlap?	Missing or inconsistent dates; control variables not used in the algorithm (e.g. breech presentation, preterm labour, polyhydramnios)	Does the output suggest data-quality problems?
Episode outcomes	Proportion of episodes with unspecific pregnancy outcome	Are clinically meaningful pregnancy endpoints being captured for each episode?	Mode of delivery	Are key delivery types captured consistently?

Classified as confidential by the European Medicines Agency

Episode-level validation – quality evaluation framework

Data source	Episode frequency (pregnancy incidence rate per 1,000 person-years in 2022)	Episode duration (% >308 days)	Episode construction (% overlap)	Episode outcomes (% without recorded pregnancy outcome)
Primary care				
BIFAP	19.8	4.1	1.7	2.6
CPRD GOLD	26.3	0.2	1.9	1.1
IPCI	29.9	1.0	2.2	46.4
SIDIAP	23.5	10.8	3.0	11.6
Secondary care				
CDW Bordeaux	30.8	1.5	0.8	2.1
FinOMOP-TaUH Pirha	39.3	27.0	4.3	1.8
IMASIS	13.0	4.5	0.4	27.4
SUCD	31.7	6.2	1.5	5.8
Claims				
InGef RDB	29.6	8.3	6.7	0.0
Registry				
DK-DHR	45.8	2.3	2.5	0.1
FinOMOP-THL	41.2	3.6	2.4	0.1
HI-SPEED	45.6	3.4	2.3	0.3
NAJS	38.6	2.6	4.7	11.5
NLHR@UiO	49.3	5.1	3.4	0.0

Classified as confidential by the European Medicines Agency

Where are we now?

- Standardised evaluation framework to assess the quality of algorithmic output (episode frequency, duration, construction, and outcomes)
- Results across 15 data sources provide a rich knowledge base for further development of the pregnancy algorithm and can inform feasibility assessment
- *PregnancyIdentifier* R package that can be used to uniformly infer pregnancy episodes across real-world data sources
- Increased depth of knowledge of pregnancy data across data sources

https://data.darwin-eu.org/EUPAS1000000812/

Classified as confidential by the European Medicines Agency

Next steps?

- Validated performance of existing algorithm at scale (across 15 data sources), improvements are possible and we are currently working on this!
- Current use: *"This early-stage package is under active development, and its performance is being optimized through ongoing evaluation. Users should carefully evaluate its suitability for their analysis and verify any outputs before using in research studies."*
  - Important to assess quality using the evaluation framework
  - Cleaning rules to restrict to clinically plausible episodes in studies

Classified as confidential by the European Medicines Agency

Thanks to all PeriNet study team members!

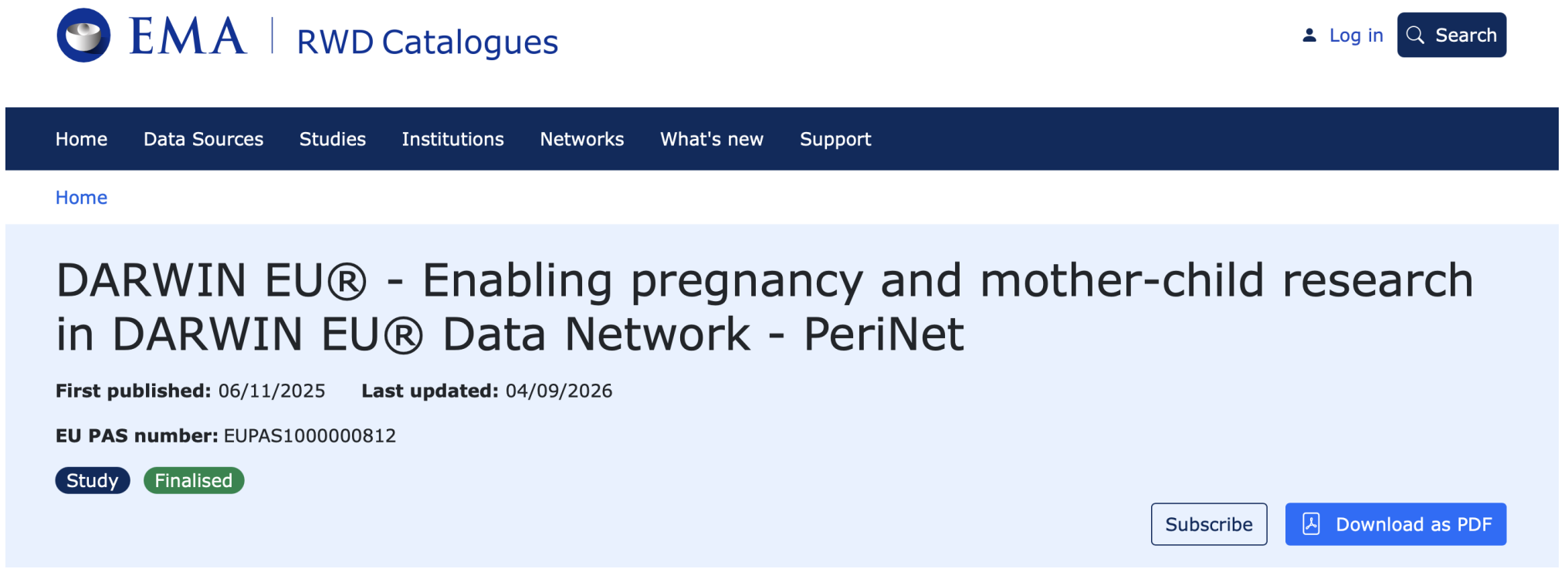
Coordinating Centre

- Julieta Politi
- Adam Black
- Aniek Markus
- Maxim Moinat
- Berta Raventós
- Maarten van Kessel
- Anum Zhar
- Ger Inberg
- Ross Williams
- Sofia Bazakou
- Katia Verhamme
- Natasha Yefimenko
- Talita Duarte-Salles

EMA

- Maria Clara Restrepo-Méndez
- Daniel Morales

Classified as confidential by the European Medicines Agency



More information: https://catalogues.ema.europa.eu/node/4716/data-management

Classified as confidential by the European Medicines Agency